2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

Performance Enhancement of Hierarchical Document Signature: A Comprehensive Study

Sukanya Manna
Research School of Computer Science
The Australian National University
Canberra ACT 0200
Email: sukanya.manna@anu.edu.au

Tom Gedeon
Research School of Computer Science
The Australian National University
Canberra ACT 0200
Email: tom.gedeon@anu.edu.au

***Abstract*—Hierarchical Document Signature (HDS) has been successfully applied in document computing to find similarity between different pieces of text [1], [2], [3]; for example sentence-sentence similarity, sentence-phrase similarity. HDS is application specific, it is dependent on different features at different levels. This paper hence presents a comprehensive study of enhancement of the performance of HDS to find semantic sentence similarity by tuning some of its significant features. The experimental results support this and show the optimal conditions at which HDS performs similarly to humans.**

***Keywords*- Fuzzy signature, Hierarchical Document Signature, sentence similarity, word similarity**

I. Introduction

It is important to abstract information available to form humanly accessible structures. The way people think and talk is hierarchical with limited information presented in any one sentence, and that information is always linked together to further information. As such, Hierarchical Document Signature (HDS) is a significant way to represent sentences when finding their similarity.

Sentences are considered to be the same if they share the same or similar meaning. There are different methods to find sentence similarities [4], [5], and [6], and HDS is one of them. Manna et al., have discussed two different HDS models in [1], [3]. In this paper we focus on the model described in [3] where similarities between sentence pairs are computed using semantic information from the source text.

The following sections briefly describe HDS, and different experiments at word and sentence level to show optimal parameters for the enhancement of performance of HDS for computing sentence similarity.

II. Brief Description of HDS

Hierarchical Document Signature (HDS) [1], [3] is a special form of Fuzzy Signature (FS) [7] meant for document analysis. In this case, the natural hierarchy of a document is maintained through the structure of HDS; document level to sentence level, sentence level to word level. HDS, like fuzzy signatures, can describe, compare and classify objects with complex structure and interdependent features. The hierarchical organization of HDS expresses the structural complexity of a problem. The local preference relations among the hierarchies and sub-branches of a document signature can be used to approximate the global preference relation of a decision problem.

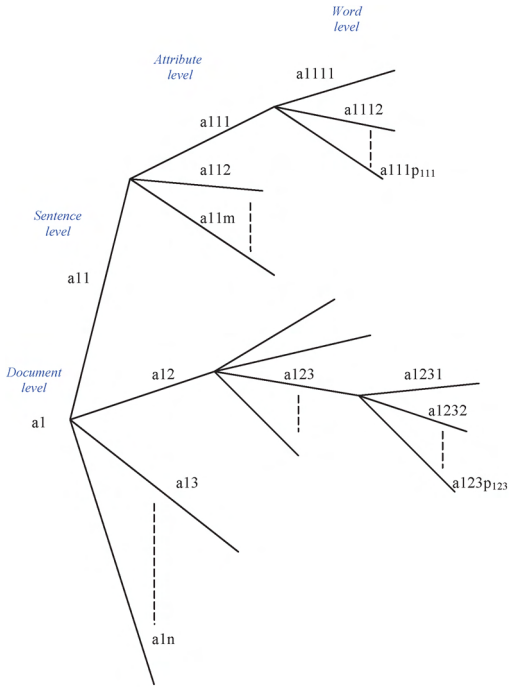

Fig. 1. Generic illustration of Hierarchical Document Signature

In fig.1, $a1$ represents a document at *document level*. A document is now segmented into n sentences, which we denote by $a11$ to $a1n$. This is the *sentence level* of the signature we developed. Next is the *attribute level*, which basically classifies the words of the sentences into their corresponding major *parts of speech*, namely *noun*, *verb*, *adjective* and *adverb*. So, for each sentence, we have m different attributes, which is fixed for each document signature. In general m is a integer, but it is constant with specific application for the ease of comparison. Now each attribute of each sentence has words, and this level is called *word level*. There can be any number of words in each attribute (POS). Here, p_{111} is the maximum number of words for attribute $a111$.

HDS can be used for computing semantic similarity of contexts. Its levels can vary depending on applications where it is used. In this case, parts of speech (POS) is a level

978-1-4244-7317-5/11/$26.00 ©2011 IEEE

which deals with the semantic information of the sentences of a document. The aggregations at different level contribute to the final similarity value. It uses Fuzzy Word Similarity (FWS) [8] at the word level to deal with the similarity or relatedness of a word pair; which then propagates to the next higher level using appropriate aggregations and continues till it reaches the document level. In this application, sentence-sentence similarity is computed.

III. Word level Scenario

The performance of HDS is mainly dependent on the word level. The similarity computed at this level is aggregated and propagated to the next higher level; thus it is essential to tune different features to achieve good performance. This word pair similarity used is known as Fuzzy Word Similarity (FWS) which was introduced in [8]. Choice of proper membership function, similarity threshold as well as the input similarity measures are different essential factors which can affect the results of FWS. So the following experiments will deal with each of these separately with proper illustrations.

a) Input Similarity Measures for FWS: As explained in [8], FWS requires two different types of similarity measures to compute PGMeasure; path based measures and gloss based measures. For path based, Lin's [9] as well as Jiang and Conrath's [10] are both experimented on. These effects can be seen in tables IV to XXVIII with different datasets in different conditions.

b) Impact of membership functions: Three sets of membership functions are shown in the table I. The membership functions are tuned based on repeated experiments and expert comments. In [8] it is seen that triangular membership functions performed better than trapezoidal at the output. But when it is further tuned, it is noticed that the trapezoidal membership functions can also be used at the output and it also gives reasonably good results. The corresponding results can be seen in tables IV to XXVIII.

TABLE I
Co-ordinates of different membership functions of fuzzy sets used here

Set I	INPUT Low_trapez	(0, 1)	(0.3, 1)	(0.4, 0)	-
	INPUT Medium_trapez	(0.3, 0)	(0.4, 1)	(0.6, 1)	(0.7, 0)
	INPUT High_trapez	(0.6, 0)	(0.8, 1)	(1, 1)	-
	OUTPUT Low_trapez	(0, 1)	(0.3, 1)	(0.4, 0)	-
	OUTPUT Medium_trapez	(0.3, 0)	(0.4, 1)	(0.6, 1)	(0.7, 0)
	OUTPUT High_trapez	(0.6, 0)	(0.8, 1)	(1, 1)	-
Set II	INPUT Low_trapez	(0, 1)	(0.3, 1)	(0.4, 0)	-
	INPUT Medium_trapez	(0.3, 0)	(0.4, 1)	(0.6, 1)	(0.7, 0)
	INPUT High_trapez	(0.6, 0)	(0.8, 1)	(1, 1)	-
	OUTPUT Low_tri	(0, 0)	(0.3, 1)	(0.6, 0)	-
	OUTPUT Medium_tri	(0.3, 0)	(0.55, 1)	(0.8, 0)	-
	OUTPUT High_tri	(0.6, 0)	(0.8, 1)	(1, 0)	-
Set III	INPUT Low_trapez	(0, 1)	(0.2, 1)	(0.3, 0)	-
	INPUT Medium_trapez	(0.2, 0)	(0.3, 1)	(0.4, 1)	(0.5, 0)
	INPUT High_trapez	(0.4, 0)	(0.5, 1)	(1, 1)	-
	OUTPUT Low_trapez	(0, 1)	(0.1, 1)	(0.3, 0)	-
	OUTPUT Medium_trapez	(0.1, 0)	(0.3, 1)	(0.5, 1)	(0.7, 0)
	OUTPUT High_trapez	(0.5, 0)	(0.7,1)	(1, 1)	-

c) Fuzzy rule-bases: In fuzzy system, there are different rule extraction methods as seen in [11] besides expert knowledge, where the rules are decided manually based on expert's understanding about the similarity measures. These rules have been used to compute FWS between a word pair.

Rule 1: If path is L AND gloss is L THEN output is L.
Rule 2: If path is L AND gloss is M THEN output is L.
Rule 3: If path is L AND gloss is H THEN output is M.
Rule 4: If path is M AND gloss is L THEN output is L.
Rule 5: If path is M AND gloss is M THEN output is M.
Rule 6: If path is M AND gloss is H THEN output is H.
Rule 7: If path is H AND gloss is L THEN output is M.
Rule 8: If path is H AND gloss is M THEN output is H.
Rule 9: If path is H AND gloss is H THEN output is H.

d) Impact of thresholds at word level: Similarity thresholds also affect the performance of HDS. Normally, for similarity computations, similarity value of 0.5 [12] is considered. But due to fuzzy approximations which initiate at word level, a new threshold of 0.4 is also experimented with and seen to be more suitable for this application.

IV. Experiments at Word Level

A. Evaluation metrics at word level

In this sub-section, evaluation of three different word similarity approaches (*path measure*, *gloss overlap* and *PGMeasure*) are presented. This has been taken directly from [8]. Unlike [12], the following metrics are defined at *word level* instead of sentence level. These are *Recall*, *Precision*, *F-measure*, *Rejection*, *Accuracy*, and *f1*. **Recall** is a proportion of correctly predicted similar word pairs compared to all similar word pairs. **Precision** is a proportion of correctly predicted similar word pairs compared to all predicted similar word pairs. **F-Measure (F1)** is a uniform harmonic mean of precision and recall. **Rejection** is a proportion of correctly predicted dissimilar word pairs compared to all dissimilar word pairs. **Accuracy** is a proportion of all correctly predicted word pairs compared to all word pairs. Lastly, **f1** is defined as a uniform harmonic mean of rejection and recall. In this work, *accuracy*, *rejection* and *f1* metrics are included in addition to the standard precision-recall based metrics as it presents another aspect of the performance based on the tradeoff between true positive and true negative judgments.

B. Evaluation process

Two words are considered to be similar if they reach certain threshold value. A scoring threshold for similar pairs is defined at 0.5 [12]. But in this case, threshold of 0.4 is also evaluated because of fuzzy approximations used in formulating PGMeasure (FWS). Thus for each kind of input and output fuzzy sets shown in table. I, results with both the thresholds are shown for each method; which will clearly explain why 0.4 has been preferred over 0.5 for word similarity computation using FWS.

C. Datasets

Results on two datasets are shown. One is data from Test of English as a Foreign Language (ESL) and the other from [13] which is called Rubenstein data in this paper. The dataset is prepared by human assessors and then experimented on.

1) ESL dataset: There is no standard benchmark dataset for computing word similarity. Generally people have evaluated their word similarities [14], [15] using some kind of multiple choice questions and their answers in TOEFL (Test of English as a Foreign Language) or ESL[1](English as a second language) or any other linguistic exams. Thus, for this work, a dataset is prepared from ESL mock exams. Like [16], the experiment is not restricted to verbs, rather all four main parts of speech are covered which have been encountered in the multiple choice questions (MCQ) in ESL. Each practice exams had several questions; with 4 choices. Out of these four, one gives you the exact match. Like Microsoft Research paraphrase corpus, [17] binary judgements are assigned to the word pair similarity instead of sentence similarity as mentioned in the dataset. For the correct choices, based on ESL answer set, 1 is assigned as relevancy score of the word pair similarity, for the rest of the incorrect answers, 0 is assigned. Antonyms are chosen deliberately for the incorrect choices (out of 3 incorrect choices) except for the questions which had some other options. Linguistically, it is difficult to determine any crisp boundary for word similarity. The binary scores are assigned as a benchmark. It is also found that for some words, which returned 'incorrect choice' in ESL test set, has some degree of similarity linguistically, might not be very relevant to the context of the question being asked. So, for these words, there are false negatives. Respective parts of speech are also assigned for the word pairs. Here, in table. II a snapshot of the dataset generated from ESL multiple choice questions is shown.

TABLE II
SNAPSHOT OF DATASET GENERATED FROM ESL MULTIPLE CHOICE QUESTIONS

Relevance	word_1	word_2	POS
1	balmy	warm	adjective
0	balmy	cold	adjective
1	pored	examined	verb
0	pored	memorized	verb
1	intersection	crossing	noun
0	intersection	ending	noun
1	repugnant	disgusting	adjective
0	repugnant	delightful	adjective
1	diligence	dedication	noun
0	diligence	ease	noun
0	widespread	limited	adjective
1	harsh	extreme	adjective
0	harsh	pleasant	adjective
1	touched	began	verb
...			

a) Comparison of different different methods: Table. III is a snapshot of the results obtained at word sense level. In some words, gloss overlap is 0, which signifies that semantically they do not share any information, but due to their taxonomical location in WordNet, they have some information in common. The PGMeasure combines these kinds of information which can be seen from this table. The tuning of membership functions can definitely provide more refined similarity values for the word pairs.

It is known that each word can have many synsets. So when similarity between two words are computed, similarity between all their combinations of synsets are computed using both gloss overlap and path measure. Maximum similarity between the synsets contribute to the word pairs' final similarity. This is used in the tables for comparison with PGMeasure.

TABLE III
WORD SIMILARITY RESULTS AT SENSE LEVEL: A SNAPSHOT OF RESULTS

(w_i#s_k,w_j#s_l)	P	G	PG
(enquiry#1,investigation#1)	0.05	0.25	0.30
(enquiry#1,investigation#2)	0.09	0.0	0.30
(enquiry#2,investigation#1)	0.48	0.0	0.30
....			
(enquiry#1,interview#1)	0.46	0.2	0.30
(enquiry#1,interview#2)	0.06	0.0	0.30
...			
(outlets#1,stores#2)	0.06	0.0	0.30
...			

b) ESL experiment results with similarity threshold = 0.5: Tables IV to XV show the comparisons of path, gloss and PGMeasure using ESL dataset with the membership functions shown in table. I at similarity threshold at 0.5 and 0.4 respectively. In this case, to measure *path measure* of two given words, maximum similarity score of the synset pairs are taken to represent the overall similarity of a word pair. Likewise for the other two measures. The effects of variation in performance due to different membership functions used, threshold, as well as different path based measures are also shown. Initially the experiments were done using Set I and Set II of table. I; where the combination of Lin's similarity score as path measure with triangular membership function for the output fuzzy set has given better performance in terms of the evaluation metrics. Inspite of 'Center Of Gravity' approximation of the trapezoidal membership functions, it is seen in that if the values of the membership functions are tuned further, then it gives very sensible results. Thus the experiments are redone using set III of table. I and the results are presented.

TABLE IV
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.2	1.0	0.33	0.61	1.0	0.33
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.24	1.0	0.39	0.63	1.0	0.39

[1]https://www.esl.org/

TABLE V
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.44	0.61	0.51	0.59	0.73	0.55
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.36	0.75	0.49	0.63	0.89	0.51

TABLE VI
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.2	1.0	0.33	0.61	1.0	0.33
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.24	1.0	0.39	0.63	1.0	0.39

TABLE VII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.44	0.61	0.51	0.59	0.73	0.55
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.36	0.75	0.49	0.63	0.88	0.51

TABLE VIII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.2	1.0	0.33	0.61	1.0	0.33
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.2	1.0	0.33	0.61	1.0	0.33

TABLE IX
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.5, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.44	0.61	0.51	0.59	0.73	0.55
Gloss	0.36	1.0	0.53	0.69	1.0	0.53
PGMeasure	0.28	0.78	0.41	0.61	0.92	0.43

Tables IV to IX illustrate the results with threshold = 0.5. For all cases, performance of PGMeasure is higher when Lin's measure is used as one of the inputs to FIS. At 0.5 threshold using Set I and II, PGMeasure performed as expected using Lin's method as a path measure. The recall is lowered but the precision and rejection has increased showing that the method could properly identify predicted similar and dissimilar word pairs using set III.

c) ESL experiment results with similarity threshold = 0.4: Tables X to XV illustrate the evaluation results with threshold of 0.4. Like the previous cases, here too using Lin's path measure as input to FIS has shown better performance than Jiang and Conrath's. With further tuning of membership functions in Set III of table. I, the recall score has improved to a greater extent. Now if the same methods are compared at threshold 0.4 and 0.5, as per expert's comments, the similarities at 0.4 level by PGMeasure are more accurate.

Using Jiang and Conrath's similarity measure as one of the inputs of FIS, the precision and rejection for all the three sets of membership functions are higher for PGMeasure than Lin's when used in FIS; but recall is lower. Recall and accuracy increased using set III of membership functions showing overall improvement with slight reduction in precision.

TABLE X
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.28	0.88	0.42	0.63	0.96	0.43
Gloss	0.36	0.9	0.51	0.67	0.96	0.52
PGMeasure	0.24	1.0	0.39	0.63	1.0	0.39

TABLE XI
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	0.55	0.51	0.55	0.62	0.54
Gloss	0.36	0.9	0.51	0.67	0.96	0.53
PGMeasure	0.36	0.69	0.47	0.61	0.85	0.51

TABLE XII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.28	0.875	0.42	0.63	0.96	0.43
Gloss	0.36	0.9	0.51	0.67	0.96	0.52
PGMeasure	0.28	1.0	0.44	0.65	1.0	0.44

2) Rubenstein dataset: Similar evaluation is done with data collected from Rubenstein's paper [13]. Words pairs are assigned a score rated by human assessors. If table. XVI is noticed, the relevance scores assigned by humans are unnormalized. But normalization is required for evaluation. So, the values are normalized by,

$$x_i^{norm} = \frac{x_i - min_X}{max_X - min_X},\ i \in [1, |X|] \quad (1)$$

where X is a set of word pairs.

TABLE XIII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	0.55	0.51	0.55	0.62	0.54
Gloss	0.36	0.9	0.51	0.67	0.96	0.52
PGMeasure	0.4	0.71	0.51	0.63	0.85	0.54

TABLE XIV
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.28	0.88	0.42	0.63	0.96	0.43
Gloss	0.36	0.9	0.51	0.67	0.96	0.52
PGMeasure	0.4	1.0	0.57	0.71	1.0	0.57

TABLE XV
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.4, ESL DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	0.55	0.51	0.55	0.62	0.54
Gloss	0.36	0.9	0.51	0.67	0.96	0.52
PGMeasure	0.64	0.70	0.67	0.69	0.73	0.68

TABLE XVI
SNAPSHOT OF RUBENSTEIN DATASET

word_1	word_2	Relevance
cord	smile	0.02
hill	woodland	1.48
rooster	voyage	0.04
car	journey	1.55
noon	string	0.04
cemetery	mound	1.69
fruit	furnace	0.05
glass	jewel	1.78
autograph	shore	0.06
magician	oracle	1.82
...		

a) Rubenstein experiment results with similarity threshold = 0.5: Now, tables XVII to XXII present evaluation result with Rubenstein's data at similarity threshold of 0.5. Like ESL's results, here too Lin's method gives better performance than Jiang and Conrath. But in this case Set III are not as good as Set I and II at this threshold.

Set I and II showed better evaluation metrics for set I and set II, but not for set III at this threshold for this dataset.

b) Rubenstein experiment results with similarity threshold = 0.4: Tables XXIII to XXVIII show performance of different similarity measures 0.4 threshold. Like the other cases illustrated so far, Lin's method for each of these sets perform better. But Set III's performance is best here and can be seen in table. XXVIII.

Now, at threshold of 0.4, using set I and set II of table. I, the evaluation metrics are little lower than 0.5. Set III has given better recall and more practical evaluation scores at 0.4 with this dataset like ESL.

TABLE XVII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.5, RUBENSTEIN DATA SET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.54	1.0	0.70	0.8	1.0	0.70
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.46	1.0	0.63	0.77	1.0	0.63

TABLE XVIII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.5, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.82	0.89	0.85	0.88	0.92	0.87
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.79	1.0	0.88	0.91	1.0	0.88

TABLE XIX
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.5, RUBENSTEIN DATA SET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.54	1.0	0.70	0.8	1.0	0.70
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.5	1.0	0.67	0.78	1.0	0.67

TABLE XX
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.5, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.82	0.89	0.85	0.88	0.92	0.87
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.79	1.0	0.88	0.91	1.0	0.88

TABLE XXI
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.5, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.54	1.0	0.70	0.8	1.0	0.70
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.36	1.0	0.53	0.72	1.0	0.53

TABLE XXII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.5, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.82	0.89	0.85	0.88	0.92	0.87
Gloss	0.36	1.0	0.53	0.72	1.0	0.53
PGMeasure	0.36	1.0	0.53	0.72	1.0	0.53

TABLE XXIII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.4, RUBENSTEIN DATA SET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	1.0	0.65	0.75	1.0	0.65
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.45	1.0	0.62	0.74	1.0	0.62

TABLE XXIV
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET I, THRESHOLD = 0.4, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.74	0.89	0.81	0.83	0.91	0.82
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.71	1.0	0.83	0.86	1.0	0.83

TABLE XXV
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.4, RUBENSTEIN DATA SET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	1.0	0.65	0.75	1.0	0.65
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.48	1.0	0.65	0.75	1.0	0.65

TABLE XXVI
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET II, THRESHOLD = 0.4, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.74	0.89	0.81	0.83	0.91	0.82
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.71	1.0	0.83	0.86	1.0	0.83

Thus from both the sets of data, it is seen that using Lin's path similarity for PGMeasure at threshold of 0.4 with Set III of table. I can give more accurate word similarity results which is more similar to human judgements.

TABLE XXVII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING JIANG AND CONRATH'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.4, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.48	1.0	0.65	0.75	1.0	0.65
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.52	1.0	0.68	0.77	1.0	0.68

TABLE XXVIII
COMPARISON OF DIFFERENT WORD SIMILARITY MEASURES USING LIN'S PATH MEASURE AS ONE OF THE INPUTS IN FIS WITH THE MEMBERSHIP FUNCTIONS OF SET III, THRESHOLD = 0.4, RUBENSTEIN DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
Path	0.74	0.89	0.81	0.83	0.91	0.82
Gloss	0.32	1.0	0.49	0.68	1.0	0.49
PGMeasure	0.74	0.89	0.81	0.83	0.91	0.82

V. SIMILARITY COMPUTATION WITH HDS

A document is modularized into sentences, and then the words of the sentences are grouped based on their POS. Using fuzzy word similarity measure (PGMeasure) similarity between two words are calculated and assigned them as the similarity score for each word in a sentences with respective POS. Then aggregations are applied at each level to get final similarity between a sentence pair. The similarity value of 0 shows the sentences are not at all related and 1 means they are identical. But practically due to fuzzy approximations at word level, the value of 1 can not be reached even if all the words are identical.

This experimental setup is taken directly from [3] with variations in the features. So, in this paper, we only mention them.

A. Evaluation Metrics at Sentence Level

Similar evaluation settings are set here as in IV-A but at sentence level. The similarities computed are evaluated with respect to the benchmark similarities provided by humans in the dataset used. The same metrics are used here as used by [12] which are namely *recall*, *precision*, *F1*, *accuracy*, *rejection*, and *f1* respectively.

The definitions are re-defined at sentence level. **Recall** is a proportion of correctly predicted similar sentences compared to all similar sentences. **Precision** is a proportion of correctly predicted similar sentences compared to all predicted similar sentences. **F1** is a uniform harmonic mean of precision and recall. **Rejection** is a proportion of correctly predicted dissimilar sentences compared to all dissimilar sentences. **Accuracy** is a proportion of all correctly predicted sentences compared to all sentences. Lastly, we define **f1** as a uniform harmonic mean of rejection and recall. A scoring threshold for positive pairs is defined at 0.4 is used unlike [12], where it was 0.5.

This has to be done due to fuzzy membership tuning at the word level.

Other than HDS, *cosine similarity*[2] is used for comparison of results; it is basically a vector space model and works on only 'bag of words' concept.

With similar reasoning as word level, in this case the experiments are done at similarity threshold of 0.4. It is also shown that set II and III of the membership functions work better for this application than set I. So we focus on experiments with these two only.

B. Dataset - The Microsoft Paraphrase Corpus

In 2005, Microsoft researchers Dolan, Brocket, and Quirck [17] published the first paraphrase corpus containing 5801 pairs of sentences with 3900 tagged as "semantically equivalent" or true paraphrases. Sentences were obtained from massive parallel news sources and tagged by 3 human raters according to guidelines described in [17]. This corpus will be referred to below as the MSRPC.

At the word level computation of HDS, membership functions mentioned in Set II and III of table. I are used by FWS. As shown in the word similarity evaluations, similarity threshold of 0.4 is used for all the evaluation in this chapter as well. Tables XXIX and XXX are based on Set II membership functions.

TABLE XXIX
EVALUATION METRIC OF SENTENCE SIMILARITY WITH 500 SENTENCE PAIRS OF MSRPC TEST DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
CosineSim	1	0.68	0.81	0.68	0	0
HDS	0.88	0.71	0.79	0.68	0.24	0.38

TABLE XXX
EVALUATION METRIC OF SENTENCE SIMILARITY WITH WHOLE MSRPC TEST DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
CosineSim	1	0.66	0.80	0.66	0	0
HDS	0.87	0.69	0.77	0.65	0.24	0.37

In table. XXX, it is seen that HDS has higher precision than cosine similarity. But there is minor drop in performance due to wrongly tagged words by the parser used. This is detected when the tagging was checked manually. There is slight rejection of 0.1 in the similarity findings. In the dataset, the humans assigned either 1 or 0 for similar or dissimilar sentences. But this is not very appropriate because similarity measurements when done by humans, can have certain level of uncertainty, which can be captured using fuzzy methods. This is actually done by HDS and the snapshot is shown in table. XXXI.

[2]http://www10.org/cdrom/papers/519/node12.html

TABLE XXXI
SNAPSHOT OF RESULTS WITH MSRPC

Sent1ID	Sent2ID	HumanScore	CosineSim	HDS Sim
3354381	3354396	0	1.0	0.37
1390995	1391183	1	0.99	0.5
2201401	2201285	0	1.0	0.55
...	...	...	...	...

c) Effect of tuning membership functions at word level computation: The HDS structure explained here is dependent on the features of the word level. The semantic similarity between words are computed by PGMeasure as discussed in [8]. The similarity values depend on the tuning of membership functions at this word level. The better the membership functions chosen, the more the results will be similar to human approximations. Here a new set (Set III of table. I) of membership functions are used and the HDS is again run on MSRP corpus and the results are observed in tables XXXII and XXXIII.

TABLE XXXII
EVALUATION METRIC OF SENTENCE SIMILARITY WITH 500 SENTENCE PAIRS OF MSRPC TEST DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
CosineSim	1.0	0.68	0.81	0.68	0	0
HDS	1.0	0.69	0.82	0.70	0.05	0.10

TABLE XXXIII
EVALUATION METRIC OF SENTENCE SIMILARITY WITH WHOLE MSRPC TEST DATASET

Methods	recall	precision	F1	accuracy	rejection	f1
CosineSim	1.0	0.66	0.80	0.66	0	0
HDS	0.99	0.67	0.80	0.67	0.03	0.06

From tables XXXII and XXXIII, it is clearly noticed that the recall and precision have improved over the previous cases.

C. Effect of different

Similarity of words are determined based on context. It is rather difficult to say which words are similar without a context. But if they belong to same parts of speech then it can be assumed that the contextual appearance of the words will also be similar if not exactly the same. Since HDS uses the word level similarity to propagate to the sentence level, it is essential to identify these words based on their proper parts of speech; as it is known that the same word in different parts of speech may have different meaning. Thus it is necessary to formulate a method which disambiguates the interpretational uncertainties and delivers more contextually appropriate word similarity - taking into account different measures, such as - graph based measures and glosses of the contextual senses. Hierarchical Document Signature (HDS) is hence chosen. It organizes a sentence into a hierarchical structure that helps to compare two or more sentences based on their part-of-speech

(POS) using fuzzy aggregation. As such, its result is effected by the result of POS tagging directly.

Now another experiment is conducted to see the effect of different POS taggers on HDS. Normally, for the previous evaluations, Stanford POS tagger is used. But it is seen that there are many instances when Stanford POS tagger produces errors and lead to wrongly tagged words. As HDS is dependent on POS, so proper tagging is required. Here, four different taggers 'hunpos' [18], 'opennlp' [19], 'crftag' [20], and 'lingpipe' [21] are used and sentence similarity is recomputed with different taggers using 500 MSRP data and the results are shown in fig. 2. It reflects that the performance of HDS can be improved using a better tagging algorithm.

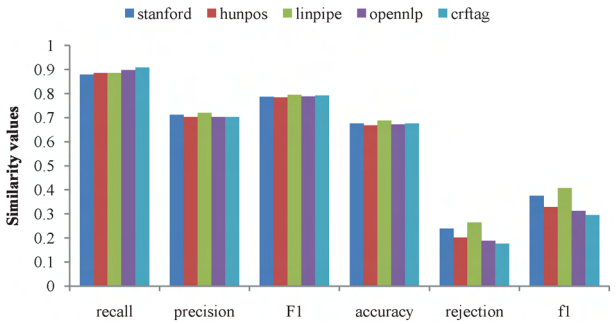

Fig. 2. Comparison of performance evaluation with varying POS taggers

VI. Conclusion

This paper discusses a detailed study of different features for improving the performance of the Hierarchical Document Signature. It is seen that proper tuning of features like fuzzy membership functions, similarity thresholds, input fuzzy sets and even a better choice of POS tagger improve HDS' performance. As future work, the main aim will be to tune the features automatically based on the nature of the dataset.

References

[1] S. Manna, B. Mendis, and T. Gedeon, "Hierarchical document signature: a specialized application of fuzzy signature for document computing," in *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*. IEEE, 2009, pp. 1083–1088.

[2] S. Manna, T. Gedeon, and B. Mendis, "Enhancement of subjective logic for semantic document analysis using hierarchical document signature," *Neural Information Processing. Theory and Algorithms*, pp. 298–306, 2010.

[3] S. Manna and T. Gedeon, "Semantic Hierarchical Document Signature for determining sentence similarity," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*. IEEE, pp. 1–8.

[4] M. Bilotti, P. Ogilvie, J. Callan, and E. Nyberg, "Structured retrieval for question answering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in information retrieval*. ACM New York, NY, USA, 2007, pp. 351–358.

[5] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2007, pp. 629–637.

[6] P. Achananuparp, X. Hu, and C. Yang, "Addressing the Variability of Natural Language Expression in Sentence Similarity with Semantic Structure of the Sentences," in *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. Springer, 2009, p. 555.

[7] B. S. U. Mendis, "Fuzzy signatures: Hierarchical fuzzy systems and alpplications (phd thesis)," Ph.D. dissertation, College of Engineering and Computer Science, The Australian National University, Australia, 2008.

[8] S. Manna and B. Mendis, "Fuzzy word similarity: A semantic approach using WordNet," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*. IEEE, pp. 1–8.

[9] D. Lin, "Using syntactic dependency as local context to resolve word sense ambiguity," in *Annual Meeting-Association For Computational Linguistics*, vol. 35. Association For Computational Linguistics, 1997, pp. 64–71.

[10] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Arxiv preprint cmp-lg/9709008*, 1997.

[11] A. Chong, T. D. Gedeon, K. W. Wong, and L. T. Kóczy, "A histogram-based rule extraction technique for fuzzy systems," in *FUZZ-IEEE*, 2001, pp. 638–641.

[12] P. Achananuparp, X. Hu, and X. Shen, "The evaluation of sentence similarity measures," in *DaWaKD08: Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*. Springer, pp. 305–316.

[13] H. Rubenstein and J. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, p. 633, 1965.

[14] T. Landauer and S. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review*, vol. 104, no. 2, pp. 211–240, 1997.

[15] D. Tatsuki, "Basic 2000 Words-Synonym Match 1," *Interactive JavaScript Quizzes for ESL Students, http://www. aitech. ac. jp/~iteslj/quizzes/js/dt/mc-2000-01syn. html*, 1998.

[16] D. Yang and D. Powers, "Word similarity on the taxonomy of WordNet."

[17] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics Morristown, NJ, USA, 2004.

[18] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: an open source trigram tagger," pp. 209–212, 2007.

[19] J. Baldridge, T. Morton, and G. Bierner, "The opennlp maximum entropy package," 2002.

[20] X. Phan, "Crftagger: Crf english pos tagger," 2006.

[21] B. Baldwin and B. Carpenter, "LingPipe."